Whole genome resequencing of a laboratory-adapted Drosophila melanogaster population sample

DATA NOTE

REVISED Whole genome resequencing of a laboratory-adapted *Drosophila melanogaster* population sample [version 3; referees: 2 approved]

William P. Gilks[1], Tanya M. Pennell[1], Ilona Flis[1], Matthew T. Webster[2], Edward H. Morrow[1]

[1]Evolution, Behaviour and Environment Group, School of Life Sciences, John Maynard Smith Building, University of Sussex, Falmer, UK
[2]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden

v3 **First published:** 07 Nov 2016, **5**:2644 (doi: 10.12688/f1000research.9912.1)
Second version: 20 Dec 2016, **5**:2644 (doi: 10.12688/f1000research.9912.2)
Latest published: 22 Dec 2016, **5**:2644 (doi: 10.12688/f1000research.9912.3)

Abstract

As part of a study into the molecular genetics of sexually dimorphic complex traits, we used high-throughput sequencing to obtain data on genomic variation in an outbred laboratory-adapted fruit fly (*Drosophila melanogaster*) population. We successfully resequenced the whole genome of 220 hemiclonal females that were heterozygous for the same Berkeley reference line genome (BDGP6/dm6), and a unique haplotype from the outbred base population (LH_M). The use of a static and known genetic background enabled us to obtain sequences from whole-genome phased haplotypes. We used a BWA-Picard-GATK pipeline for mapping sequence reads to the dm6 reference genome assembly, at a median depth-of coverage of 31X, and have made the resulting data publicly-available in the NCBI Short Read Archive (Accession number SRP058502). We used Haplotype Caller to discover and genotype 1,726,931 small genomic variants (SNPs and indels, <200bp). Additionally we detected and genotyped 167 large structural variants (1-100Kb in size) using GenomeStrip/2.0. Sequence and genotype data are publicly-available at the corresponding NCBI databases: Short Read Archive, dbSNP and dbVar (BioProject PRJNA282591). We have also released the unfiltered genotype data, and the code and logs for data processing and summary statistics (https://zenodo.org/communities/sussex_drosophila_sequencing/).

Open Peer Review

Referee Status: ✔ ✔

	Invited Referees 1	2
REVISED version 3 published 22 Dec 2016		
REVISED version 2 published 20 Dec 2016		
version 1 published 07 Nov 2016	✔ report	✔ report

1 **Stephen Richards**, Baylor College of Medicine USA

2 **Geraldine A. Van der Auwera**, Broad Institute of Harvard and MIT USA

Discuss this article

Comments (0)

Corresponding authors: William P. Gilks (wpgilks@gmail.com), Edward H. Morrow (ted.morrow@sussex.ac.uk)

How to cite this article: Gilks WP, Pennell TM, Flis I *et al.* **Whole genome resequencing of a laboratory-adapted *Drosophila melanogaster* population sample [version 3; referees: 2 approved]** *F1000Research* 2016, **5**:2644 (doi: 10.12688/f1000research.9912.3)

Grant information: Funding was provided to EM by a Royal Society University Research Fellowship, the Swedish Research Council (No. 2011-3701), and by the European Research Council (No. 280632).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 07 Nov 2016, **5**:2644 (doi: 10.12688/f1000research.9912.1)

REVISED **Amendments from Version 2**

In this version, a mistake in the title of the Version 2 has been corrected - the words 'population sample' have been added to the title.

See referee reports

Introduction

As part of a study on the molecular genetics of sexually dimorphic complex traits, we used hemiclonal analysis in conjunction with high-throughput sequencing[1] to characterise molecular genetic variation across the genome, from an outbred laboratory-adapted population of *Drosophila melanogaster*, LH_M[2,3]. The hemiclone experimental design allows the repeated phenotyping of multiple individuals, each with the same unrecombined haplotype on a different random genetic background. This method has been used to investigate standing genetic variation and intersexual genetic correlations for quantitative traits[2] and gene expression[4], but it has not yet been used to obtain genomic data.

The 220 hemiclone females that were sequenced in the present study have a maternal haplotype, from the *dm6* reference assembly strain (BDGP6+ISO1 mito/dm6, Bloomington *Drosophila* Stock Center no. 2057)[5,6], and have a different paternal genome each, sampled using cytogenetic cloning from the LH_M base population (See Figure 1). All non-reference genotypes in the sequenced LH_M hemiclones were expected to be heterozygous and in-phase, except in rare instances where the in-house dm6 reference strain also had the same non-reference allele.

Previous studies indicate that the limits for DNA quantity in next-generation sequencing are 50–500ng[7]. We sequenced individual *D. melanogaster*, rather than pools of clones, because more biological information can be obtained, and because modern transposon-based library preparation allows accurate sequencing at low concentrations of DNA. *D. melanogaster* is a small insect (~1μg) although this problem is off-set by the reduced proportion of repetitive intergenic sequence, and small genome size relative to other insects (170Mb verses ~500Mb)[7].

We mapped reads to the *D. melanogaster* dm6 reference assembly using a BWA-Picard-GATK pipeline, and called nucleotide variants using both HaplotypeCaller, and Genomestrip, the latter of which detects copy-number variation up to 1Mb in length. A graphic representation of the data analysis pipeline is provided in Figure 2. We have made the mapped sequencing data, and genotype data publicly-available on NCBI, and additionally have made the metadata, analysis code and logs publicly-available on Zenodo. This is the first report of a study which uses methods for detecting both SNPs, indels and structural variants (deletions and duplications >1Kb in length), genome-wide in next-generation sequencing data, and the first report of whole genome resequencing in hemiclonal individuals.

Materials and methods

Study samples

The base population (LH_M) was originally established from a set of 400 inseminated females, trapped by Larry Harshman in a citrus orchard near Escalon, California in 1991[3]. It was initially kept at a large size (more than 1,800 reproducing adults) in the lab of William Rice (University College Santa Barbara, USA). In 1995 (approximately 100 generations since establishment) the rearing protocol was changed to include non-overlapping generations, and a moderate rearing density with 16 adult pairs per vial (56 vials in total) during 2 days of adult competition, and 150–200 larvae during the larval competition stage[3]. In 2005, a copy of LH_M population sample was transferred to Uppsala University, Sweden (approximately 370 generations since establishment), and in 2012, to the University of Sussex (UK), when the current set of 223 haplotypes were sampled. At the point of sampling we estimate that the population had undergone 545 generations under laboratory conditions, 445 of which had been using the same rearing protocol.

Hemiclonal lines were established by mating groups of five clone-generator females (C(1)DX,*y*,*f*; T(2;3) *rdgC st in ri* p^P bw^D) with 230 individual males sampled from the LH_M base population (see 2). A single male from each cross was then mated again to a group of five clone-generator females in order to amplify the number of individuals harbouring the sampled haplotype. Seven lines failed to become established at this point. The remaining 223 lines were maintained in groups of up to sixteen stock hemiclonal males in two vials that were transferred to fresh vials each week. Stock hemiclonal males were replenished every six weeks by mating with groups of clone-generator females. A stock of reference genome flies (Bloomington *Drosophila* Stock Center no. 2057) was established and maintained initially using five rounds of sib-sib matings before expansion. 223 virgin reference genome females were then collected and mated to a single male from each of 223 hemiclonal lines. Female offspring from this cross therefore have one copy of the reference genome and one copy of the hemiclonal haplotype. Groups of these hemiclonal females were collected as virgins, placed in 99% ethanol and stored at -20°C prior to DNA extraction.

DNA extraction

One virgin female per hemiclonal line, was homogenised with a microtube pestle, followed by 30-minute mild-shaking incubation in proteinase K. DNA was purified using the DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA), according to manufacturer's instructions. Volumes were scaled-down according to mass of input material. Barrier pipette tips were used throughout, in order to minimise cross-contamination of DNA. Template assessment using the Qubit BR assay (Thermo Fischer, NY, USA) indicated double-stranded DNA, 10.4Kb in length at concentrations of 2–4 ng/μl (total quantity 50–100ng).

Whole-genome resequencing

Sequencing was performed under contract by Exeter Sequencing service, University of Exeter, UK. The sonication protocol for shearing of the DNA was optimised for low concentrations to generate fragments 200–500bp in length. Libraries were prepared and indexed using the Nextera Library Prep Kit (Illumina, San Diego,

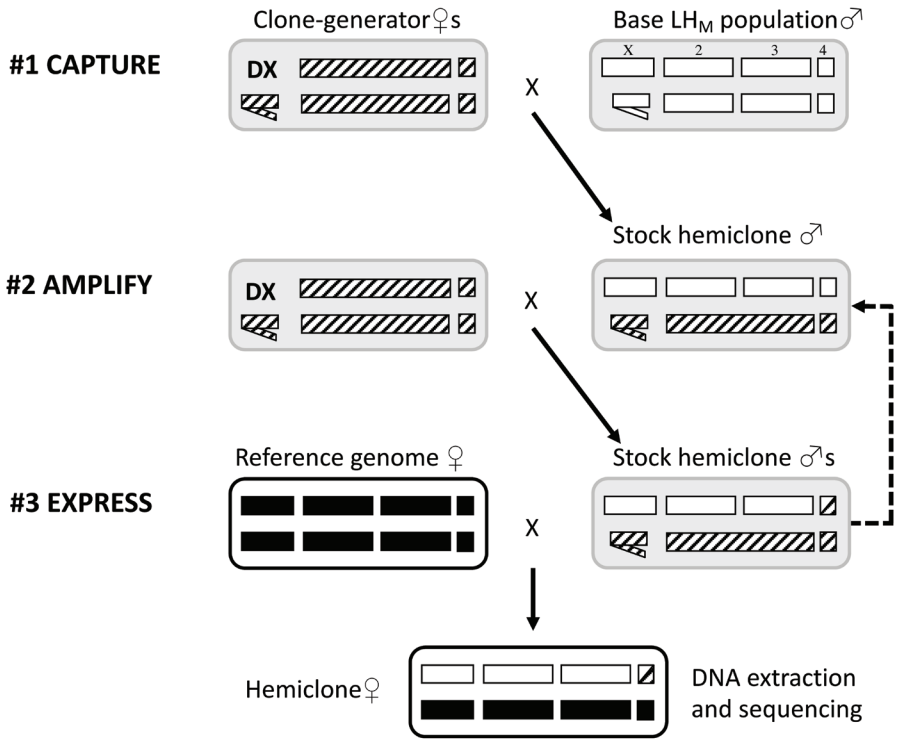

Figure 1. Breeding design for generating each hemiclonal line. 1 Capture - Single wild-type male from the base-population (open chromosomes) was crossed to five clone generator female, which harbour a fused double-X chromosome (DX), a Y chromosome, as well as a marked translocation of chromosomes 2 and 3 (hatched chromosomes; see text for full genotype). This cross captured a single wild-type haplotype. 2 . Amplify - A single heterozygous male was then crossed again to a group of clone-generator females to amplify the number stock hemiclonal males. This cross can be repeated, to replenish the stock hemiclone males. 3 Express - Finally, a single reference genome female (filled chromosomes) was crossed to a single stock hemiclonal male to produce a female that harbours the original wild-type haplotype (excluding the 4th dot chromosome, which remains uncontrolled throughout), in a reference genome genetic background. Cytoplasmic factors in the offspring also derive from the reference genome stock (black outline). DNA is extracted from a single hemiclone female from each line and sequenced.

USA). All samples were sequenced on a HiSeq 2500 (Illumina), with five individuals per lane. We also sequenced DNA from two individuals from the in-house reference line (Bloomington *Drosophila* Stock Centre no. 2057). One was prepared as the hemiclones, using the Illumina Nextera library (sample RGil), and the other using an older, Illumina Nextflex method (sample RGfi). The median number of read pairs across all samples was 29.23×10^6 (IQR 14.07×10^6). Quality metrics for the sequencing data were generated with FastQC v0.10.0 by Exeter Biosciences, and used to determine whether results were suitable for further analyses. For twelve samples with less than 8×10^6 reads, sequencing was repeated successfully (H006, H041, H061, H084, H086, H087, H092, H098, H105), with a further three samples omitted entirely (H015, H016, H136), leaving 220 hemiclonal samples in total. As shown in Figures 3A and 3B, the read quality score and quality-per-base for the samples taken forward for genotyping in this study were well within acceptable standards, and similar across all samples.

Read mapping

Raw data (*fastq* files) were stored and processed in the Linux Sun Grid Engine in the High-Performance Computing facility, University of Sussex. Adaptor sequences (Illumina Nextera N501-H508 and N701-N712), poor quality reads (Phred score <7) and short reads were removed using Fastq-mcf (ea-utils v.1.1.2). Settings were: log-adapter minimum-length-match: 2.2, occurrence threshold before adapter clipping: 0.25, maximum adapter difference: 10%, minimum remaining length: 19, skew percentage-less-than causing cycle removal: 2, bad reads causing cycle

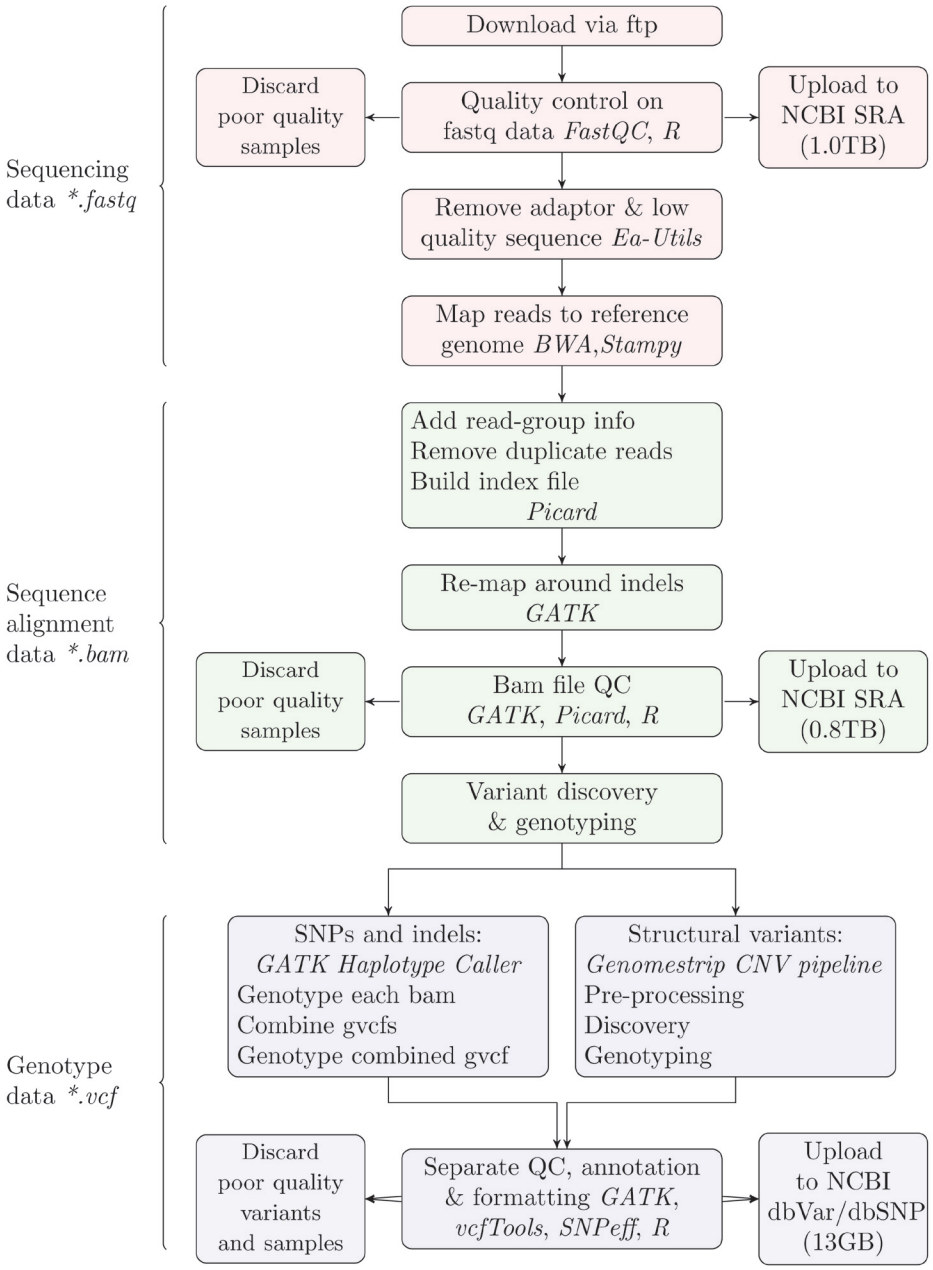

Figure 2. Flow diagram for the high-throughput sequencing data analysis. Data file types are indicate at the left side. Nodes are also coloured by file type. Software platforms are in typed in italics. Values in parentheses indicate the size of the data uploaded at different stages. Latex code for generating this figure is available at https://doi.org/10.5281/zenodo.168582.

removal: 20%, quality threshold causing base removal: 10, window-size for quality trimming: 1, number of reads to use for sub-sampling: 3×10^5.

Cleaned sequence reads were mapped to the *D. melanogaster* genome assembly, release 6.0 (Assembly Accession GCA_000001215.4[6]) using Burrows-Wheeler Aligner *mem* (version 0.7.7-r441)[8], with a mapping quality score threshold of 20. Remaining reads were remapped using Stampy v1.0.24, which is slower but more precisely maps reads which are divergent from the reference genome assembly[9]. This method was used previously for the *Drosophila* Genome Nexus[10]. Removal of duplicate reads, indexing and sorting was performed with Picard-Tools v1.77. Re-mapping of sequence reads around insertion-deletion polymorphisms was performed using Genome Analysis Tool-Kit (GATK) v3.2.2, as a recommended standard practice[11].

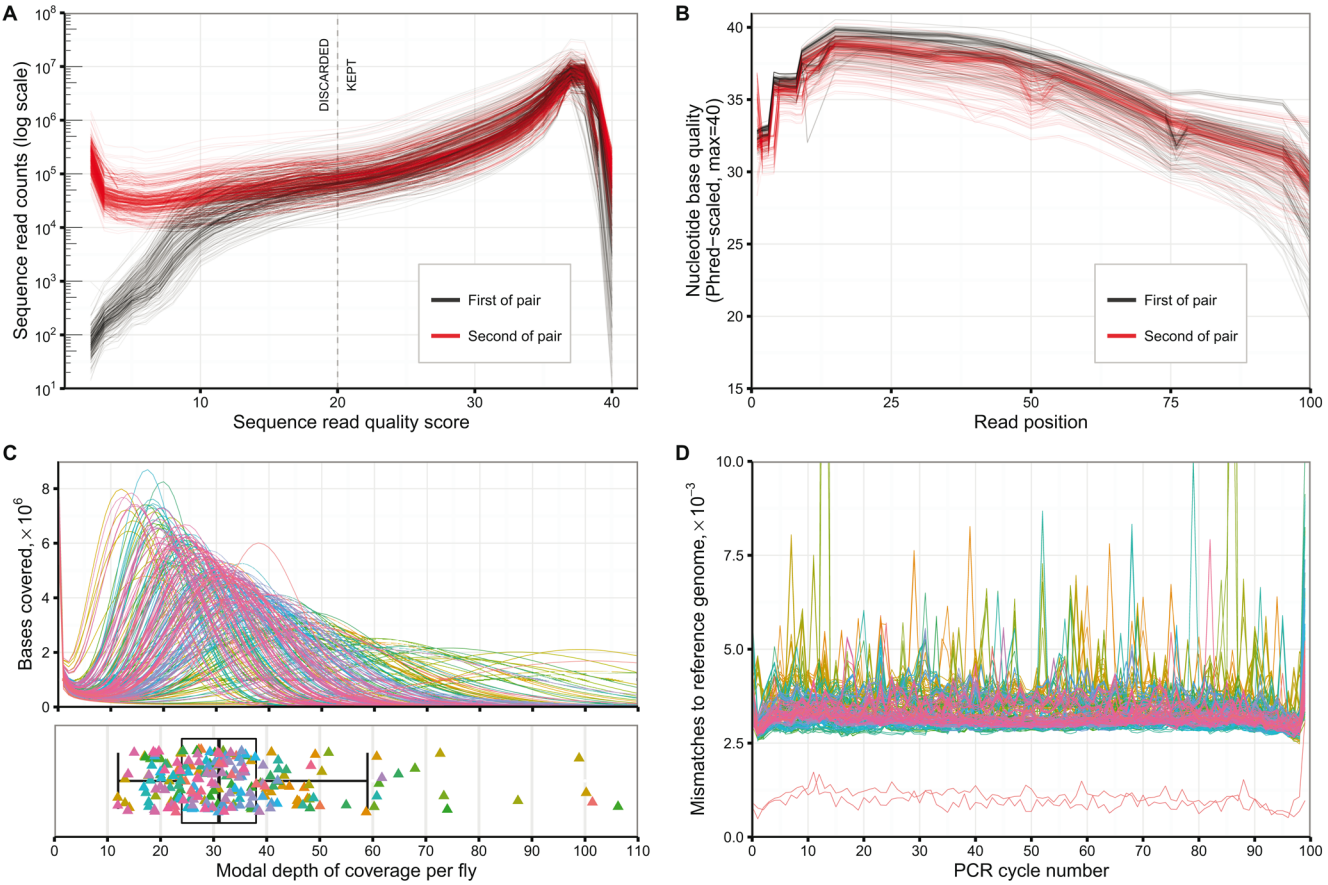

Figure 3. Next-generation sequencing assessment. A: Sequence read quality for each sample sequenced. Y-axis scale is logarithmic. **B**: Quality of sequences by nucleotide base position for each sample. **C**: Read depth of coverage distribution across each sample. Colouring corresponds to the order which which the samples were originally sequenced. **D**: Mis-matches to the dm6 reference genome assembly, by PCR cycle-number. Colouring is by sample as in plot **C**. The two red lines with visibly-lower mismatch rates than the others correspond to the two in-house BDGP/dm6 reference lines that were sequenced. Data and R code for this figure are located at https://doi.org/10.5281/zenodo.159282.

The median depth of coverage across all samples used for genotyping was 31X (IQR 14, see Figure 3C). As shown in Figure 3D, the mean nucleotide mis-match rate to the dm6 reference assembly for the LH_M hemiclones was 3.27×10^{-3} per PCR cycle (IQR 0.2×10^{-3}), contrasting with the two reference line samples for which the mis-match rate was $0.89\text{-}1.10\times10^{-3}$ per cycle. We observed spikes of nucleotide mismatches in some PCR cycles for some samples, which are likely to be errors rather than true sequence variation.

Small-variant detection methods

Single-nucleotide polymorphisms (SNPs) and insertion/deletions (indels) <200bp in length, were detected and genotyped relative to the BDGP+ISO1/dm6 assembly, on chromosomes 2,3,4,X, and mitochondrial genome using Haplotyper Caller (GATK v3.4-0)[12]. Individual bam files were genotyped, omitting reads with a mapping quality under 20, stand call and emit confidence thresholds of 31, then combined and genotyped again. 143,726,002 bases of genomic sequence were analysed from which 1,996,556 variant loci were identified consisting of 1,581,341 SNPs, 196,582 deletions, and 218,633 insertions. Functional annotation was added using SNPeff v4.1[13].

We used hard-filtering to remove variants generated by error, because the alternative 'variant recalibration' requires prior information on variant positions from a similar population or parents. Quality filtering thresholds were decided following inspection of the various sequencing metrics associated with each variant locus, and by software developers' recommendations[12]. The filtering thresholds were: Quality-by-depth >2, strand bias <50 (Phred-scaled *p*-value from Fisher's Exact test), mapping quality >58, mapping quality rank sum >-7.0, read position rank sum >-5.0, combined read depth <15000, and call rate >90%. This filtering removed 167,319 variants (8.3%), leaving 1,829,237. Summary values for the variant quality metrics are shown in Table 1. Distributions of quality metrics for Haplotype Caller variants are shown in Supplementary Figure 2. The density of sequence variants, measured as the median for windows of 10Kb in length across the genome, was 75 per for biallelic SNPs, 1 for multi-allelic SNPs, 6 for bi-allelic indels, and

Table 1. Haplotype Caller variant quality metrics and genotype frequencies.

Variant type N	SNPs (biallelic) 1,411,395	SNPs (multi) 43,798	Indels (biallelic) 138,687	Indels (multi) 65,660
Total depth	6440 (1725)	6316 (2100)	6134 (1836)	5973 (2081)
Event length	0 (0)	0 (0)	2 (5)	1 (8)
Strand bias*	1.12 (2.25)	1.34 (3.14)	1.76 (3.88)	1.77 (4.45)
Mapping quality	62.12 (6.18)	64.94 (8.57)	71.17 (12.77)	69.58 (11.36)
Map qual rank sum	0.25 (1.04)	0.9 (2.37)	3.14 (3.21)	2.68 (2.91)
Quality-by-depth	16.65 (3.51)	17(3.81)	18.52 (6.21)	16.96 (6.39)
Quality	34968 (62236)	57028 (67558)	25842 (59889)	40479 (63590)
Genotype counts				
Reference	151 (120)	102(122)	166(114)	122(123)
Heterozygous	70 (118)	117(121)	54(114)	95(122)
Homozygous non-ref.	0 (0)	0(0)	0(0)	0(0)
No call	0 (1)	1(4)	0(2)	2(5)

Values show the total number of variants, median (and IQR) for each metric. Data generated from *vcf* file using GATK VariantsToTable, on the quality-filtered data. *Strand bias refers to Phred-scaled *p*-value from Fisher's Exact Test. Code and data used to generate this table located at https://doi.org/10.5281/zenodo.159282.

3 for multi-allelic indels (see Figure 4A). Mean separation between variants of any type or allele frequency was 78bp. As shown in Figure 4B the allele frequency distribution for bi-allelic SNPs and indels was similar, and broadly within expectations for an out-bred diploid population sample. The two in-house reference line individuals had 515 homozygous and 3171 heterozygous mutations from the reference assembly. The median genotype counts for the 220 LH_M hemiclone individuals, were 585 homozygous, 728,214 heterozygous and 4963 no-call (IQR 400, 36707 and 7876). Genotype counts for each individual are shown in Figure 4C.

For data submission to NCBI dbSNP, we were obliged to exclude 44,644 indels that were multi-allelic or greater than 50bp in length, and a further 57,662 SNPs and indels situated within deletions. Variants greater than 50bp in length were submitted to the NCBI structural variant database dbVar. The genotype data submitted to dbSNP consists of 1,726,931 quality-filtered, functionally-annotated variant records (1,423,039 SNPs and 303,892 short, biallelic insertion and deletion variants) corresponding to 383,378,682 individual genotype calls.

Structural-variant detection methods

Large genomic variants – deletions and duplications, between 1Kb and 100Kb in length – were detected and genotyped using GenomeStrip v2.0[14]. One of the reference strain individuals (sample RGfi) was omitted from the this analysis because a different sequencing library preparation method was used from the other samples (see above). We included the following settings (according to developers' guidelines): Sex-chromosome and k-mer masking when estimating sequencing depth, computation of GC-profiles and read counts, and reduced insert size distributions. Large variant discovery and genotyping was performed only on chromosomes 2, 3, 4 and X, omitting the mitochondrial genome and unmapped scaffolds.

We used the Genomestrip CNV Discovery pipeline with the settings: minimum refined length 500, tiling window size 1000, tiling window overlap 500, maximum reference gap length 1000, boundary precision 100, and genotyped the results with the GenerateHaploidGenotypes R script (genotype likelihood threshold 0.001, R version 3.0.2). Following visualisation of the genotype results and comparison with the *bam* sequence alignment files using the Integrated Genomics Viewer (IGV) v2.3.72[15], we excluded telomeric and centromeric regions where the sequencing coverage was fragmented, and six regions of multi-allelic gains of copy-number with dispersed breakpoints, previously reported to undergo mosaic *in vivo* amplification prior to oviposition[16] (see Supplementary Table 1 for genomic positions, and Supplementary Figure 3 for visualisation of *in vivo* amplification in a sequence alignment file). We excluded 6 samples (H082, H083, H090, H097, H098, H153) for which 80–90% of the genome was reported by Genomestrip to contain structural variation, which we regarded as error. Most these samples were grouped by the order in which they were processed for DNA extraction and sequencing, so this may have been caused partly by a batch-effect leading to differences in read pair separation, depth-of-coverage, and response to normal fluctuations in GC-content. Following removal of these samples, there were 2897 CNVs (1687 deletions, 877 duplications, and 333 of the 'mixed' type), ranging in size from 1000bp to 217,707bp. We observed eight regions, for which Genomestrip identified multiple adjacent CNVs in single individuals, but which are likely single CNVs, 100Kb to 1.3Mb in length (Supplementary Table 2).

Quality-filtering for structural variants detected by Genomestrip analysis of whole-genome resequencing data are not thoroughly established. We visually inspected, in the *bam* read alignment files using the Integrated Genomics Viewer[15], reported structural variants which were most likely to be artefacts. Specifically these were variants with: i) Extreme values for quality-score, GC-content or cluster separation, ii) Any homozygous non-reference genotypes (not expected with our breeding design), iii) Type 'mixed'. Following this, we used the following criteria for quality filtering: Quality score >15, cluster separation <17, GC-fraction >0.33, no mixed types (deletions and duplications only), homozygous non-reference genotype count >0, and heterozygous genotype count <200. Summaries of the quality metrics for quality-filtered data are shown in Table 2 and Supplementary Figure 2. We applied an upper limit to the cluster separation to remove groups of outliers in the upper end of the distribution, although this may have

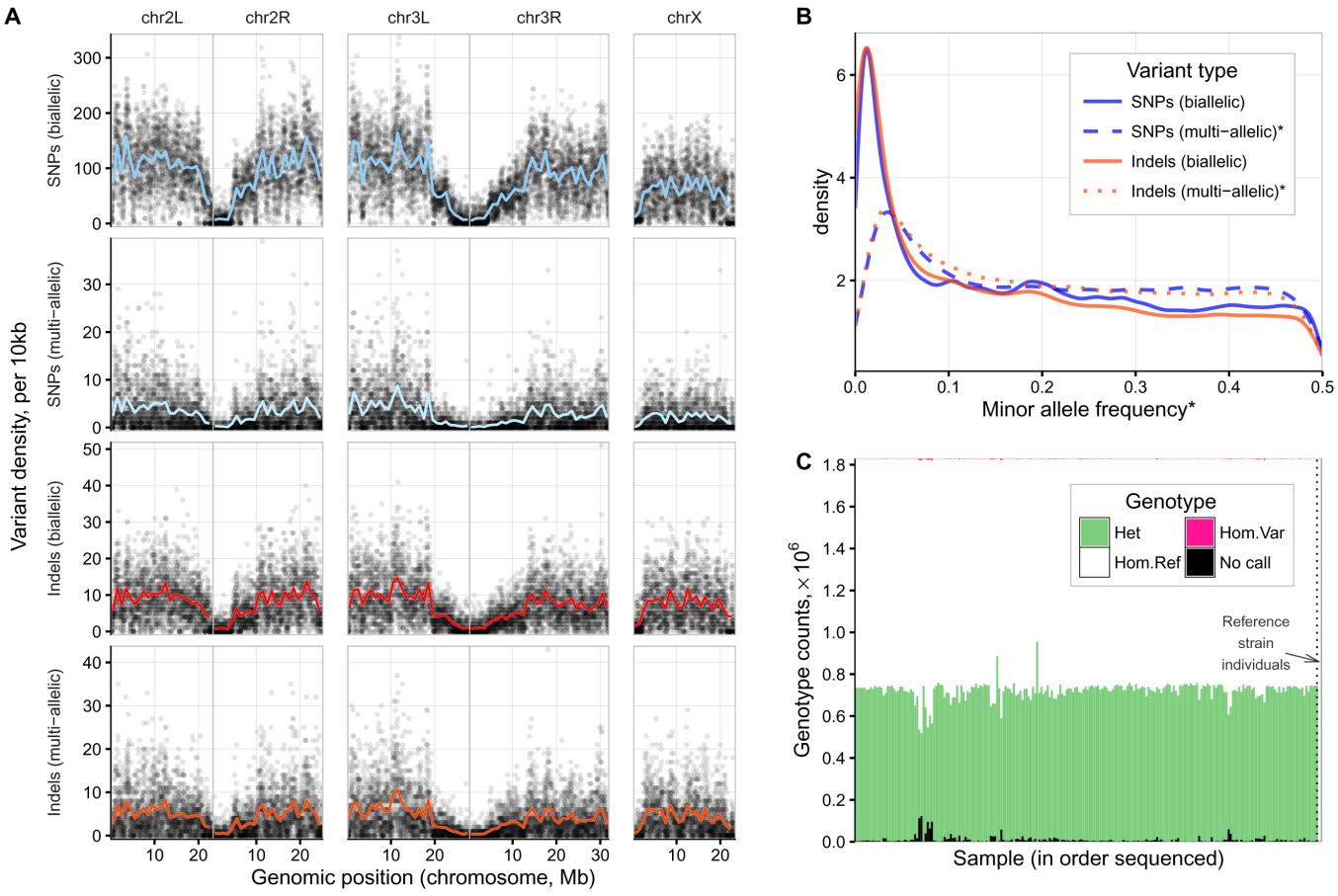

Figure 4. Haplotype Caller small variant results. A: Density of common variants across the genome (MAF>0.05 (Variants from the in-house reference line are included but account for less than 3,686 of the 1,825,917 common variants plotted (<0.2%). **B**: Allele frequency distribution by variant type. *MAF values were calculated from the count of heterozygous calls, and so for multi-allelic variants, the MAF is derived from the combined count of both alternate alleles. **C**: Genotype counts per individual genotyped. Data generated using GATK/3.4 VariantEvaluation function. Data and R code for this figure are located at https://doi.org/10.5281/zenodo.159282.

Table 2. Quality metrics for Genomestrip CNVs.

Metric N	Deletions 78	Duplications 89
GC-fraction	0.39 (0.07)	0.42 (0.06)
Cluster separation	8.84 (3.70)	9.78 (3.17)
Quality	103.93 (505.71)	490.95 (1128.32)
Heterozygote count (max 213)*	22.00 (42.50)	42.00 (53.00)
Length (kb)	2.20 (3.54)	3.40 (2.35)

Values show the total number of variants, median (and IQR) for each metric. Data generated from *vcf* file using GATK VariantsToTable, on the quality-filtered data. *No CNVs in the quality-filtered samples had a 'no-call' or homozygous non-reference genotype.

excluded many true, low-frequency variants. However, data on rare variants are not directly useful for our further investigations.

After filtering, 167 CNVs remained (78 deletions and 89 duplications, size range 1Kb-26.6Kb). The positions and genotypes of these CNVs for each individual are shown in Figure 5. The genotype data for quality-filtered CNVs were combined with the data from 2252 indels >50bp from the Haplotype Caller pipeline, and a total of 2419 variants were uploaded to the public database on structural variation, NCBI dbVar. Although we have used methods for detecting SNPs, indels and CNVs, variants between 200bp and 1Kb are not reported by either HaplotypeCaller or Genomestrip. Additionally, sequence inversions are not detected by these methods and the upper limits to CNV detection using Genomestrip, based on the parameters and results of this study, are 100Kb-1Mb.

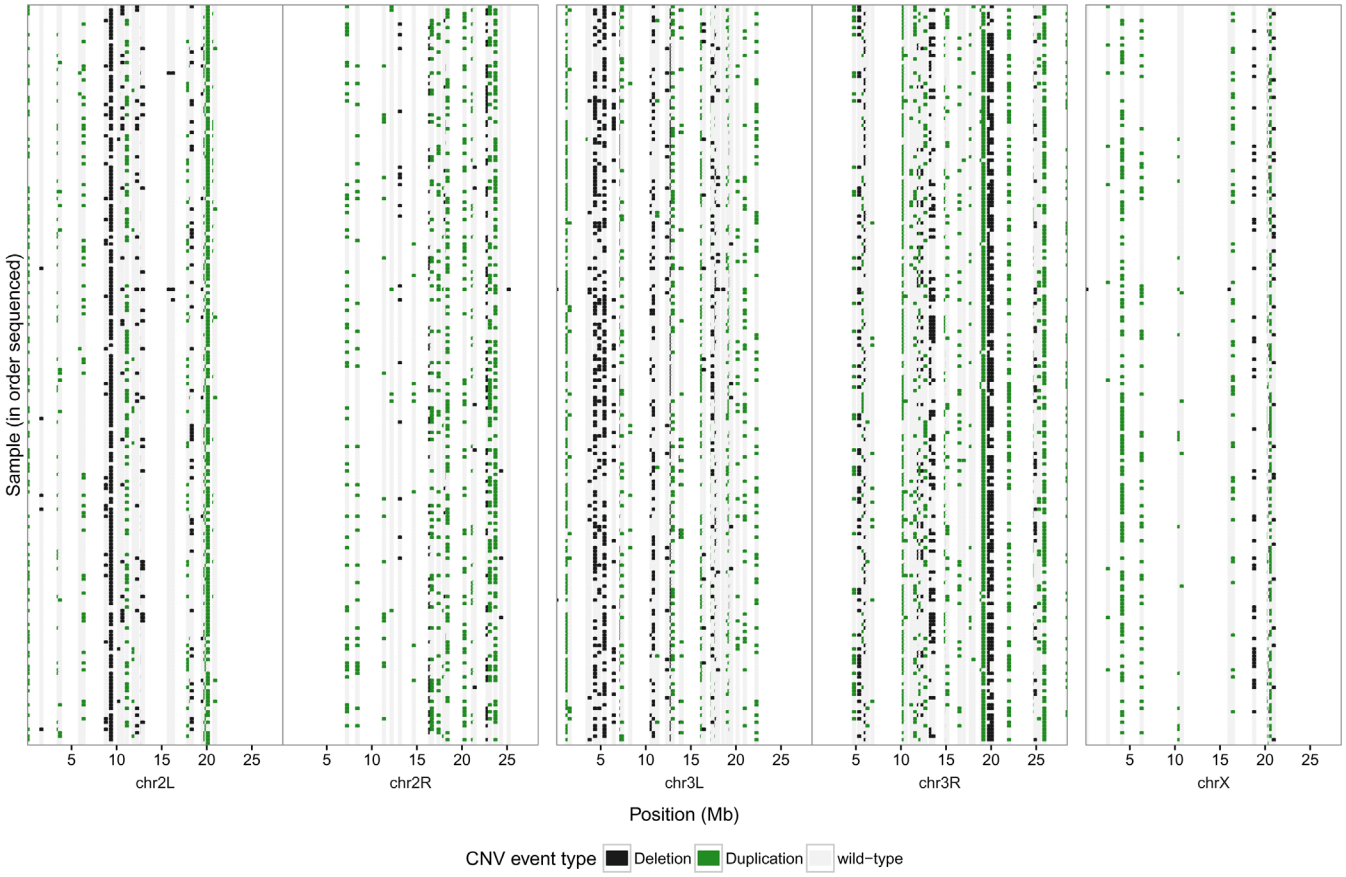

Figure 5. Genomestrip structural variant results across the *D. melanogaster* genome. Each row corresponds to an individual sequenced (in order originally sequenced from top to bottom, with the reference line at the bottom). Image generated using R/3.3.1 (package ggplot v2.1.0) with data generated by GATK VariantsToTable with individual genotypes as copy-numbers. Data and R code for this figure are located at https://doi.org/10.5281/zenodo.159282.

Dataset validation

Initial validation of our methods can be seen by lack of variants in the two reference line individuals compared with the LH_M hemiclones (3,686 verses a median of 728,799 per sample). For a more thorough test of the genotyping and hemiclone method reproducibility, we sequenced an additional hemiclone individual from three of the LH_M lines, and mapped the reads to the reference genome assembly as before. For HaplotypeCaller, we generated gVCF files for each sample, and then performed genotyping and quality-filtering as described above, except that the original three samples were replaced with the replication test samples. Similarly, for Genomestrip, we performed structural variant discovery and genotyping on all of the same samples as before, replacing three original samples with the replication test samples. We then used the GATK Genotype Concordance function to generate counts of genotype differences between the three pairs of samples. Overall results are presented in Table 3. Genotype reproducibility for quality-filtered biallelic SNPs was 98.5–99.5%, going down to 89.1–93.2% for filtered multi-allelic indels. Reproducibility of structural variant genotype calls was 95.6–100.0%, although we noted that for one individual (H119) filtering actually reduced the reproducibility rate from 99.7% to 95.6%. Full code, logs and numerical results can be found at http://doi.org/10.5281/zenodo.160539.

Although these results indicate that our genotype accuracy is very good, there are several caveats to consider. In the quality-filtered small-variant data, seven samples (H034, H035, H040, H038, H039, H188, H174) had prominently higher genotype drop-out rates than the others (of 2–7%), as well as a higher proportion of homozygous non-reference genotypes (2–4%; See Figure 4C). Additionally two samples had prominently more heterozygous variants (H072:885,551 and H093:955,148 verses the other LH_M hemiclones: mean 710,934).

Although the genotype replication rate for the structural variants was also very high, we cannot exclude the possibility that, due to incomplete masking of hard-to-sequence regions of the reference assembly, variants which are artefacts reported in the original genotype data, may also be present in the replication genotype data.

Data availability

All publicly-available records are for 220 LH_M hemiclone individuals and 2 in-house reference line individuals, with the exception of the large-variant data for which one in-house reference line sample and six LH_M hemiclones were omitted. The NCBI BioProject identifier is PRJNA282591. Code, logs and quality control data for each dataset, and for generating the figures and tables in this

Table 3. Genotype reproducibility rates(%)*.

Variant type	Sample ID	Unfiltered	Filtered
HaplotypeCaller/3.4			
Bi-allelic SNP	H119	98.9	99.5
	H137	97.7	98.5
	H151	97.8	98.3
Multi-allelic SNP	H119	95.0	96.6
	H137	92.3	94.0
	H151	92.1	93.6
Bi-allelic indel	H119	98.1	98.6
	H137	96.3	96.8
	H151	96.0	96.4
Multi-allelic indel	H119	91.9	93.2
	H137	88.0	89.3
	H151	87.9	89.1
Genomestrip/2.0			
Deletion	H119	99.7	95.6
	H137	100.0	100.0
	H151	100.0	100.0
Duplication	H119	99.7	100.0
	H137	99.9	100.0
	H151	99.6	100.0

*Presented values are the overall genotype concordance, as generated using GATK/3.4 Genotype Concordance function. Code, logs and output data are available at http://doi.org/10.5281/zenodo.160539.

manuscript are publicly-available at Zenodo, https://zenodo.org/, 'Sussex Drosophila Sequencing' community. Use of the files uploaded to Zenodo is under Creative Commons 4.0 license.

Sequencing data

Raw *fastq* sequence reads, and *bam* alignment files for *D. melanogaster* are publicly-available at the NCBI Sequence Read Archive, accession number SRP058502 (https://www.ncbi.nlm.nih.gov/sra/?term=SRP058502). The code for read-mapping, alongside the run logs and quality-control data are available at https://doi.org/10.5281/zenodo.159251. Additionally the sequence alignment files for the corresponding *Wolbachia* have accession number SRP091004 (https://www.ncbi.nlm.nih.gov/sra/?term=SRP091004), with further supporting files at https://doi.org/10.5281/zenodo.159784[16].

Small-variant data

Records of quality-filtered sequence variants identified by GATK HaplotypeCaller in the LH_M hemiclones, and in the in-house reference line, are available from the NCBI dbSNP, https://www.ncbi.nlm.nih.gov/projects/SNP/snp_viewBatch.cgi?sbid=1062461, handle: MORROW_EBE_SUSSEX. In compliance with NCBI dbSNP criteria, variants >50bp in length, multi-allelic indels, and variants with a null alternate allele are excluded. More extensive genotype data (unfiltered, quality-filtered, and formatted for NCBI dbSNP) are available at https://doi.org/10.5281/zenodo.159272[17]. Also included is the code used for variant discovery and genotyping, quality-filtering and formatting, alongside run logs and quality-control data. Further filtering of this dataset may be necessary to remove localised areas of artefact SNPs in single samples. We have also released a gvcf genotypes file which contains an 'all-sites' record of the sample genotypes, available for download at https://doi.org/10.5281/zenodo.198880.

Structural-variant data

Records of quality-filtered variants detected by GenomeStrip, and variants >50bp detected by Haplotype Caller are publicly-available at NCBI dbVar, accession number nstd134, https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd134/. Unfiltered and filtered genotype data, code for CNV discovery and genotyping using Genomestrip/2.0, run logs, and summary data are publicly-available at https://doi.org/10.5281/zenodo.159472[17].

Genotype reproducibility testing

Run code and logs for performing the genotyping using Haplotype Caller and Genomestrip when three samples are replaced by hemiclones from the same line, code for comparing the genotype calls between pairs of hemiclones, and results tables are located at https://doi.org/10.5281/zenodo.160539.

Data for manuscript tables and figures

Input data, code and logs for generating the figures and tables used in this manuscript are located at https://doi.org/10.5281/zenodo.159282. The LATEXcode for generating the flow-diagram for Figure 2 is available from https://doi.org/10.5281/zenodo.168582. Code and logs for the generation of the input data is provided in the data releases pertaining to each process.

Author contributions

EM conceived and supervised the experiment. EM, TP, IF, MW and WG designed the experiment. TP and IF established and

maintained the lines, and carried out the DNA extractions. WG analysed the sequencing and genotype data. WG and MW developed the read-mapping and variant-calling procedures. WG and EM wrote the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

Funding was provided to EM by a Royal Society University Research Fellowship, the Swedish Research Council (No. 2011-3701), and by the European Research Council (No. 280632).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

Sequencing was performed under contract by Exeter University, DNA Sequencing Service (UK), who also provided analysis advice, http://www.exeter.ac.uk/business/facilities/sequencing/. Crucial computational support was provided by Jeremy Maris at the Centre for High-Performance Computing, University of Sussex, http://www.sussex.ac.uk/its/services/research/highperformance. Bob Handsaker (Harvard Medical School, USA) provided analysis advice for use of Genomestrip for structural variant detection.

Supplementary material

Supplementary information for: "Whole genome resequencing of a laboratory-adapted *Drosophila melanogaster* population".

Click here to access the data

References

1. Bentley DR, Balasubramanian S, Swerdlow HP, *et al.*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature.* 2008; **456**(7218): 53–59.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Abbott JK, Morrow EH: **Obtaining snapshots of genetic variation using hemiclonal analysis.** *Trends Ecol Evol.* 2011; **26**(7): 359–368.
PubMed Abstract | Publisher Full Text
3. Rice WR, Linder JE, Friberg U, *et al.*: **Inter-locus antagonistic coevolution as an engine of speciation: assessment with hemiclonal analysis.** *Proc Natl Acad Sci U S A.* 2005; **102**(Suppl 1): 6527–6534.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Innocenti P, Morrow EH: **The sexually antagonistic genes of *Drosophila melanogaster*.** *PLoS Biol.* 2010; **8**(3): e1000335.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Adams MD, Celniker SE, Holt RA, *et al.*: **The genome sequence of *Drosophila melanogaster*.** *Science.* 2000; **287**(5461): 2185–2195.
PubMed Abstract | Publisher Full Text
6. Hoskins RA, Carlson JW, Wan KH, *et al.*: **The Release 6 reference sequence of the *Drosophila melanogaster* genome.** *Genome Res.* 2015; **25**(3): 445–458.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Richards S, Murali SC: **Best Practices in Insect Genome Sequencing: What Works and What Doesn't.** *Curr Opin Insect Sci.* 2015; **7**: 1–7.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–2079.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Lunter G, Goodson M: **Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads.** *Genome Res.* 2011; **21**(6): 936–939.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Lack JB, Cardeno CM, Crepeau MW, *et al.*: **The *Drosophila* genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population.** *Genetics.* 2015; **199**(4): 1229–1241.
PubMed Abstract | Publisher Full Text | Free Full Text
11. DePristo MA, Banks E, Poplin R, *et al.*: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet.* 2011; **43**(5): 491–498.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Van der Auwera GA, Carneiro MO, Hartl C, *et al.*: **From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.** *Curr Protoc Bioinformatics.* 2013; **11**(1110): 11.10.1–11.10.33.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Cingolani P, Platts A, Wang le L, *et al.*: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain *w*^{1118}; *iso*-2; *iso*-3.** *Fly (Austin).* 2012; **6**(2): 80–92.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Handsaker RE, Van Doren V, Berman JR, *et al.*: **Large multiallelic copy number variations in humans.** *Nat Genet.* 2015; **47**(3): 296–303.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Robinson JT, Thorvaldsdóttir H, Winckler W, *et al.*: **Integrative genomics viewer.** *Nat Biotechnol.* 2011; **29**(1): 24–26.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Spradling AC, Mahowald AP: **Amplification of genes for chorion proteins during oogenesis in *Drosophila melanogaster*.** *Proc Natl Acad Sci U S A.* 1980; **77**(2): 1096–1100.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Gilks W: **Read-mapping for next-generation sequencing data (Wolbachia) [Data set].** *Zenodo.* 2016.
Data Source
18. Gilks W: **SNP and indel discovery and genotyping in next-generation sequencing data [Data set].** *Zenodo.* 2016.
Data Source

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 16 November 2016

doi:10.5256/f1000research.10683.r17452

✔ **Geraldine A. Van der Auwera**
Data Science and Data Engineering group, Data Sciences Platform, Broad Institute of Harvard and MIT, Cambridge, MA, USA

Overall this is a very solid technical note. The methods seem sound, the descriptions are fairly straightforward, and caveats are properly acknowledged. The authors have provided detailed descriptions of what was done (including software tool versions) and made data and code available to reproduce not only the dataset but also all figures in the paper itself.

Regarding the experimental design, I think it's great that Gilks et al. chose to perform sequencing on individual flies rather than pooled samples. It takes some extra effort to deal with the small amounts of starting material involved, but the resulting dataset is that much more valuable.

It's also nice to see a study looking at CNVs and short variants together. As tools in this space improve and enable greater integration I look forward to seeing more analysis of how the different variant types relate to each other (e.g. looking at which short variants might be amplified or suppressed by CNV events).

Request for additional figures

I would recommend including diagrams of the hemiclonal experimental design and of the analysis workflows to maximize clarity. In particular, I think it could be made more obvious that the HaplotypeCaller workflow was run using the GVCF pathway for joint analysis.

Minor comments

I prefer "high-throughput sequencing" to "next generation sequencing" (this technology was "next-gen" ten years ago, now it's just the current standard).

On page 3, does "Fine mapping" refer to realignment around indels or equivalent processes?

On page 4, I would express "strand bias" as "FisherStrand estimation of strand bias" to avoid ambiguity with other estimators like Strand Odds Ratio, SOR).

On page 4, does "null alternate alleles" refer to the GATK convention of emitting "*" to record sites with spanning deletions, as documented here: https://software.broadinstitute.org/gatk/guide/article?id=6926?

Typos (Page/paragraph)

P2p3 - "detections" -> "detects"
P2p6 - "off-spring" -> "offspring"
P3p1 - "were were" -> "were"
P4p1 - "mis-matches" -> "mismatches"
P6p2 - "g.vcf" -> "GVCF" or "gVCF"

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: Not really a competing interest as such, but I'd like to disclose that I am a member of the group that develops the GATK software so I have strong opinions about how it should be run.

Author Response 12 Dec 2016

William Gilks, University of Sussex, UK

Dear Dr Van der Auwera,

We thank you for reviewing our manuscript, and consider that your suggestions improve the clarity and quality of the manuscript, particularly for technical accuracy, and overall communication. Following these suggestions, we describe the changes that we have made:

1. As requested, we have included a figure describing the breeding of the hemiclonal lines and indicating what is happening to the chromosomes, (Figure1), including a reference to the figure in the main text.

2. As requested, we have included figure that summarises the analysis pipeline for the next-generation sequencing data and genotyping procedures (Figure 2).

3. In accordance with the suggestion, we have changed the term “next-generation sequencing” to "high-throughput sequencing". We have also added a reference to for the sequencing method (Introduction, first sentence, Bently et al 2008 Nature PMID:18987734).

4. We agree that our description of the sequence-read mapping procedures was generally vague and inaccurate (Page 3, Read Mapping methods, 2nd paragraph).

   Our original sentences were: “Fine mapping was performed with both Stampy v1.0.248 and the Genome Analysis Tool-Kit (GATK) v3.2.29 (following10). Removal of duplicate reads, indexing and sorting was performed with Picard- Tools v1.77 and SamTools v1.0.”

   We have changed this to: “Remaining reads were re-mapped using Stampy v1.0.24, which is slower but more precisely maps reads which are divergent from the reference genome assembly9. This method was used previously for the Drosophila Genome Nexus10. Removal of duplicate reads, indexing and sorting was performed with Picard-Tools v1.77. Re-mapping of sequence reads around insertion-deletion polymorphisms was performed using Genome Analysis Tool-Kit (GATK) v3.2.2, as a recommended standard practice11.”

   The updated text provides more information on the properties of secondary mapping using Stampy, and how it has been used previously for the Drosophila Genome Nexus.

Furthermore, the new text, distinguishes the process of fine-mapping of reads around insertion-deletion polymorphisms using GATK.

5. You suggested for clarification, changing "strand bias" to "FisherStrand estimation of strand bias" in order to avoid ambiguity with other estimators (page 4, Small variant detection methods, 2nd paragraph, 3rd sentence). We have added in brackets a definition for strand bias, in this case as 'Phred-scaled p-value from Fisher's Exact test'. We have also added this information to Table 1, on HaplotypeCaller variant metrics.

6. On page 4 (Small variant detection methods, last paragraph, preparation of data to NCBI dbSNP), you query whether our use of the term 'null alternate allele' as a reference to the GATK convention of using an asterisk symbol for an alternate allele which is located in a spanning deletion. We have removed the term 'null alternate allele', and merely stated that variants located within deletions were excluded.

   The original sentence was: "For data submission to dbSNP, we removed 44,644 indels that were multi-allelic or greater than 50bp in length, and a further 57,662 variants that had null alternate alleles (likely due to being situated within a deletion)."

   The new sentence is: "For data submission to NCBI dbSNP, we were obliged to exclude 44,644 indels that were multi-allelic or greater than 50bp in length, and a further 57,662 SNPs and indels situated within deletions."

7. We have made the grammatical and spelling corrections as suggested.

We hope that these alterations meet your approval, and would be happy to make any further changes that may be required.

Sincerely,

William Gilks
Tanya Pennell
Ilona Flis
Matthew Webster
Edward Morrow

Competing Interests: No competing interests to declare.

Referee Report 08 November 2016

doi:10.5256/f1000research.10683.r17453

Stephen Richards
Human Genome Sequencing Center, Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, TX, USA

As a data note I think this is an excellent very detailed and comprehensive description of a dataset.
I can find all the data in the public databases as described.

I have only the most minor of quibbles:

1. The breeding of the hemiclonal lines always confuses me, and I think it would help the reader if there were a figure describing this with different colored chromosomes showing what is happening as you go through the crosses.

2. If I wanted a vcf file (or ideally gvcf) for the project is there one available for download.

3. Maybe stick the data in fly-var? http://www.iipl.fudan.edu.cn/FlyVar/

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 12 Dec 2016

William Gilks, University of Sussex, UK

Dear Dr Richards,

We thank you for reviewing our manuscript, and consider that your suggestions improve the quality of the manuscript, and the public availability of data. Following your suggestions, we have made the following changes:

1. As requested, we have included a figure describing the breeding design of the hemiclonal lines, which hopefully clarifies the transmission of the different chromosomes in each cross (Figure 1).

2. You suggested releasing a 'gvcf' file for public use, which contains a record of the genotype information at all sites in the D. melanogaster genome in the LHM population study sample. We had previously deleted this file after the smaller vcf file was generated. In response to the suggestion we have re-generated a gvcf, and deposited it in Zenodo (https://doi.org/10.5281/zenodo.198880), alongside code and run-logs, and made a note of this in the manuscript text under 'Data Availability; Small variant data'. This gvcf differs from the original gvcf only in that: i) An updated version of GATK was used (3.4 compared to 3.2), and ii) That all scaffolds of the dm6 assembly were analysed, including those which have not been mapped to specific chromosomal positions.

3. You suggested that we deposit the genotype data in the 'Fly-var' database (http://www.iipl.fudan.edu.cn/FlyVar/). Following communication with the curators of Fly-var, we understand that the current method of data submission is by e-mail, which is unsuitable for our large dataset. We have given the URLs for our data to the Fly-var curators ready for upload when procedures exist.

We hope that these alterations meet your approval, and would be happy to make any further changes that may be required.

Sincerely,

William Gilks

Tanya Pennell
Ilona Flis
Matthew Webster
Ted Morrow

Competing Interests: No competing interests to declare.